Using Linguistic Annotations in Statistical Machine Translation of Film Subtitles

Using Linguistic Annotations in Statistical Machine Translation of Film Subtitles

Christian Hardmeier
Fondazione Bruno Kessler
Human Language Technologies
Via Sommarive, 18
I-38050 Povo (Trento)
hardmeier@fbk.eu

Martin Volk
Universität Zürich
Inst. für Computerlinguistik
Binzmühlestrasse 14
CH-8050 Zürich
volk@cl.uzh.ch

Abstract

Statistical Machine Translation (SMT) has been successfully employed to support translation of film subtitles. We explore the integration of Constraint Grammar corpus annotations into a Swedish–Danish subtitle SMT system in the framework of factored SMT. While the usefulness of the annotations is limited with large amounts of parallel data, we show that linguistic annotations can increase the gains in translation quality when monolingual data in the target language is added to an SMT system based on a small parallel corpus.

1 Introduction

In countries where foreign-language films and series on television are routinely subtitled rather than dubbed, there is a considerable demand for efficiently produced subtitle translations. Although superficially it may seem that subtitles are not appropriate for automatic processing as a result of their literary character, it turns out that their typical text structure, characterised by brevity and syntactic simplicity, and the immense text volumes processed daily by specialised subtitling companies make it possible to produce raw translations of film subtitles with statistical methods quite effectively. If these raw translations are subsequently post-edited by skilled staff, production quality translations can be obtained with considerably less effort than if the subtitles were translated by human translators with no computer assistance.

A successful subtitle Machine Translation system for the language pair Swedish–Danish, which has now entered into productive use, has been presented by Volk and Harder (2007). The goal of the present study is to explore whether and how the quality of a Statistical Machine Translation (SMT) system of film subtitles can be improved by using linguistic annotations. To this end, a subset of 1 million subtitles of the training corpus used by Volk and Harder was morphologically annotated with the DanGram parser (Bick, 2001). We integrated the annotations into the translation process using the methods of factored Statistical Machine Translation (Koehn and Hoang, 2007) implemented in the widely used Moses software. After describing the corpus data and giving a short overview over the methods used, we present a number of experiments comparing different factored SMT setups. The experiments are then replicated with reduced training corpora which contain only part of the available training data. These series of experiments provide insights about the impact of corpus size on the effectivity of using linguistic abstractions for SMT.

2 Machine translation of subtitles

As a text genre, subtitles play a curious role in a complex environment of different media and modalities. They depend on the medium film, which combines a visual channel with an auditive component composed of spoken language and non-linguistic elements such as noise or music. Within this framework, they render the spoken dialogue into written text, are blended in with the visual channel and displayed simultaneously as the original sound track is played back, which redundantly contains the same information in a form that may or may not be accessible to the viewer. In their linguistic form, subtitles should be faithful, both in contents and in style, to the film dialogue which they represent. This means in particular that they usually try to convey an impression of orality. On the other hand, they are constrained by the mode of their presentation: short, written captions superimposed on the picture frame.

According to Becquemont (1996), the characteristics of subtitles are governed by the interplay of two conflicting principles: *unobtrusiveness* (discrétion) and *readability* (lisibilité). In

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 57–64

order to provide a satisfactory experience to the viewers, it is paramount that the subtitles help them quickly understand the meaning of the dialogue without distracting them from enjoying the film. The amount of text that can be displayed at one time is limited by the area of the screen that may be covered by subtitles (usually no more than two lines) and by the minimum time the subtitle must remain on screen to ensure that it can actually be read. As a result, the subtitle text must be shortened with respect to the full dialogue text in the actors' script. The extent of the reduction depends on the script and on the exact limitations imposed for a specific subtitling task, but may amount to as much as 30 % and reach 50 % in extreme cases (Tomaszkiewicz, 1993, 6).

As a result of this processing and the considerations underlying it, subtitles have a number of properties that make them especially well suited for Statistical Machine Translation. Owing to their presentational constraints, they mainly consist of comparatively short and simple phrases. Current SMT systems, when trained on a sufficient amount of data, have reliable ways of handling word translation and local structure. By contrast, they are still fairly weak at modelling long-range dependencies and reordering. Compared to other text genres, this weakness is less of an issue in the Statistical Machine Translation of subtitles thanks to their brevity and simple structure. Indeed, half of the subtitles in the Swedish part of our parallel training corpus are no more than 11 tokens long, including two tokens to mark the beginning and the end of the segment and counting every punctuation mark as a separate token. A considerable number of subtitles only contains one or two words, besides punctuation, often consisting entirely of a few words of affirmation, negation or abuse. These subtitles can easily be translated by an SMT system that has seen similar examples before.

The orientation of the genre towards spoken language also has some disadvantages for Machine Translation systems. It is possible that the language of the subtitles, influenced by characteristics of speech, contains unexpected features such as stutterings, word repetitions or renderings of non-standard pronunciations that confuse the system. Such features are occasionally employed by subtitlers to lend additional colour to the text, but as they are in stark conflict with the ideals of unobtrusiveness and readability, they are not very frequent.

It is worth noting that, unlike rule-based Machine Translation systems, a statistical system does not in general have any difficulties translating ungrammatical or fragmentary input: phrase-based SMT, operating entirely on the level of words and word sequences, does not require the input to be amenable to any particular kind of linguistic analysis such as parsing. Whilst this approach makes it difficult to handle some linguistic challenges such as long-distance dependencies, it has the advantage of making the system more robust to unexpected input, which is more important for subtitles.

We have only been able to sketch the characteristics of the subtitle text genre in this paper. Díaz-Cintas and Remael (2007) provide a detailed introduction, including the linguistics of subtitling and translation issues, and Pedersen (2007) discusses the peculiarities of subtitling in Scandinavia.

3 Constraint Grammar annotations

To explore the potential of linguistically annotated data, our complete subtitle corpus, both in Danish and in Swedish, was linguistically analysed with the DanGram Constraint Grammar (CG) parser (Bick, 2001), a system originally developed for the analysis of Danish for which there is also a Swedish grammar. Constraint Grammar (Karlsson, 1990) is a formalism for natural language parsing. Conceptually, a CG parser first produces possible analyses for each word by considering its morphological features and then applies constraining rules to filter out analyses that do not fit into the context. Thus, the word forms are gradually disambiguated, until only one analysis remains; multiple analyses may be retained if the sentence is ambiguous.

The annotations produced by the DanGram parser were output as tags attached to individual words as in the following example:

```
$-
Vad   [vad] <interr> INDP NEU S NOM @ACC>
vet   [veta] <mv> V PR AKT @FS-QUE
du    [du] PERS 2S UTR S NOM @<SUBJ
om    [om] PRP @<PIV
det   [den] <dem> PERS NEU 3S ACC @P<
$?
```

In addition to the word forms and the accompanying lemmas (in square brackets), the annotations

contained part-of-speech (POS) tags such as INDP for "independent pronoun" or V for "verb", a morphological analysis for each word (such as NEU S NOM for "neuter singular nominative") and a tag specifying the syntactic function of the word in the sentence (such as @ACC>, indicating that the sentence-initial pronoun is an accusative object of the following verb). For some words, more fine-grained part-of-speech information was specified in angle brackets, such as <interr> for "interrogative pronoun" or <mv> for "verb of movement". In our experiments, we used word forms, lemmas, POS tags and morphological analyses. The fine-grained POS tags and the syntax tags were not used.

4 Factored Statistical Machine Translation

Statistical Machine Translation formalises the translation process by modelling the probabilities of target language (TL) output strings T given a source language (SL) input string S, $p(T \mid S)$, and conducting a search for the output string $\hat{T}$ with the highest probability. In the Moses decoder (Koehn et al., 2007), which we used in our experiments, this probability is decomposed into a log-linear combination of a number of feature functions $h_i(S,T)$, which map a pair of a source and a target language element to a score based on different submodels such as translation models or language models. Each feature function is associated with a weight λ_i that specifies its contribution to the overall score:

$$\begin{aligned}\hat{T} &= \arg\max_T \log p(T \mid S) \\ &= \arg\max_T \sum_i \lambda_i h_i(S,T)\end{aligned}$$

The *translation models* employed in factored SMT are phrase-based. The phrases included in a translation model are extracted from a word-aligned parallel corpus with the techniques described by Koehn et al. (2003). The associated probabilities are estimated by the relative frequencies of the extracted phrase pairs in the same corpus. For *language modelling*, we used the SRILM toolkit (Stolcke, 2002); unless otherwise specified, 6-gram language models with modified Kneser-Ney smoothing were used.

The SMT decoder tries to translate the words and phrases of the source language sentence in the order in which they occur in the input. If the target language requires a different word order, reordering is possible at the cost of a score penalty. The translation model has no notion of sequence, so it cannot control reordering. The language model can, but it has no access to the source language text, so it considers word order only from the point of view of TL grammaticality and cannot model systematic differences in word order between two languages. *Lexical reordering models* (Koehn et al., 2005) address this issue in a more explicit way by modelling the probability of certain changes in word order, such as swapping words, conditioned on the source and target language phrase pair that is being processed.

In its basic form, Statistical Machine Translation treats word tokens as atomic and does not permit further decomposition or access to single features of the words. Factored SMT (Koehn and Hoang, 2007) extends this model by representing words as vectors composed of a number of features and makes it possible to integrate word-level annotations such as those produced by a Constraint Grammar parser into the translation process. The individual components of the feature vectors are called *factors*. In order to map between different factors on the target language side, the Moses decoder works with *generation models*, which are implemented as dictionaries and extracted from the target-language side of the training corpus. They can be used, e. g., to generate word forms from lemmas and morphology tags, or to transform word forms into part-of-speech tags, which could then be checked using a language model.

5 Experiments with the full corpus

We ran three series of experiments to study the effects of different SMT system setups on translation quality with three different configurations of training corpus sizes. For each condition, several Statistical Machine Translation systems were trained and evaluated.

In the *full data* condition, the complete system was trained on a parallel corpus of some 900,000 subtitles with source language Swedish and target language Danish, corresponding to around 10 million tokens in each language. The feature weights were optimised using minimum error rate training (Och, 2003) on a development set of 1,000 subtitles that had not been used for training, then the system was evaluated on a 10,000 subtitle test

set that had been held out during the whole development phase. The translations were evaluated with the widely used BLEU and NIST scores (Papineni et al., 2002; Doddington, 2002). The outcomes of different experiments were compared with a randomisation-based hypothesis test (Cohen, 1995, 165–177). The test was two-sided, and the confidence level was fixed at 95 %.

The results of the experiments can be found in table 1. The baseline system used only a translation model operating on word forms and a 6-gram language model on word forms. This is a standard setup for an unfactored SMT system. Two systems additionally included a 6-gram language model operating on part-of-speech tags and a 5-gram language model operating on morphology tags, respectively. The annotation factors required by these language models were produced from the word forms by suitable generation models.

In the *full data* condition, both the part-of-speech and the morphology language model brought a slight, but statistically significant gain in terms of BLEU scores, which indicates that abstract information about grammar can in some cases help the SMT system choose the right words. The improvement is small; indeed, it is not reflected in the NIST scores, but some beneficial effects of the additional language models can be observed in the individual output sentences.

One thing that can be achieved by taking word class information into account is the disambiguation of ambiguous word forms. Consider the following example:

Input: Ingen vill bo mitt emot en ismaskin.
Reference: Ingen vil bo lige over for en ismaskine.
Baseline: Ingen vil bo mit imod en ismaskin.
POS/Morphology: Ingen vil bo over for en ismaskin.

Since the word *ismaskin* ‘ice machine’ does not occur in the Swedish part of the training corpus, none of the SMT systems was able to translate it. All of them copied the Swedish input word literally to the output, which is a mistake that cannot be fixed by a language model. However, there is a clear difference in the translation of the phrase *mitt emot* ‘opposite’. For some reason, the baseline system chose to translate the two words separately and mistakenly interpreted the adverb *mitt*, which is part of the Swedish expression, as the homonymous first person neuter possessive pronoun ‘my’, translating the Swedish phrase as ungrammatical Danish *mit imod* ‘my against’. Both of the additional language models helped to rule out this error and correctly translate *mitt emot* as *over for*, yielding a much better translation. Neither of them output the adverb *lige* ‘just’ found in the reference translation, for which there is no explicit equivalent in the input sentence.

In the next example, the POS and the morphology language model produced different output:

Input: Dåliga kontrakt, dålig ledning, dåliga agenter.
Reference: Dårlige kontrakter, dårlig styring, dårlige agenter.
Baseline: Dårlige kontrakt, dårlig forbindelse, dårlige agenter.
POS: Dårlige kontrakt, dårlig ledelse, dårlige agenter.
Morphology: Dårlige kontrakter, dårlig forbindelse, dårlige agenter.

In Swedish, the indefinite singular and plural forms of the word *kontrakt* ‘contract(s)’ are homonymous. The two SMT systems without support for morphological analysis incorrectly produced the singular form of the noun in Danish. The morphology language model recognised that the plural adjective *dårlige* ‘bad’ is more likely to be followed by a plural noun and preferred the correct Danish plural form *kontrakter* ‘contracts’. The different translations of the word *ledning* as ‘management’ or ‘connection’ can be pinned down to a subtle influence of the generation model probability estimates. They illustrate how sensitive the system output is in the face of true ambiguity. None of the systems presented here has the capability of reliably choosing the right word based on the context in this case.

In three experiments, the baseline configuration was extended by adding lexical reordering models conditioned on word forms, lemmas and part-of-speech tags, respectively. As in the language model experiments, the required annotation factors on the TL side were produced by generation models.

The lexical reordering models turn out to be useful in the *full data* experiments only when conditioned on word forms. When conditioned on lemmas, the score is not significantly different from the baseline score, and when conditioned on part-of-speech tags, it is significantly lower. In this case, the most valuable information for lexical reordering lies in the word form itself. Lemma and part of speech are obviously not the right abstractions to model the reordering processes when sufficient data is available.

Table 1 Experimental results

	full data		*symmetric*		*asymmetric*	
	BLEU	NIST	BLEU	NIST	BLEU	NIST
Baseline	53.67 %	8.18	42.12 %	6.83	44.85 %	7.10
Language models						
parts of speech	⋆ 53.90 %	8.17	⋆ 42.59 %	6.87	○ 44.71 %	7.08
morphology	⋆ 54.07 %	8.18	⋆ 42.86 %	6.92	⋆ 44.95 %	7.09
Lexical reordering						
word forms	⋆ 53.99 %	8.21	42.13 %	6.83	○ 44.72 %	7.05
lemmas	53.59 %	8.15	⋆ 42.30 %	6.86	○ 44.71 %	7.06
parts of speech	○ 53.36 %	8.13	⋆ 42.33 %	6.86	○ 44.63 %	7.05
Analytical translation	53.73 %	8.18	⋆ 42.28 %	6.90	⋆ 46.73 %	7.34

⋆ BLEU score significantly above baseline ($p < .05$)
○ BLEU score significantly below baseline ($p < .05$)

Another system, which we call the *analytical translation* system, was modelled on suggestions by Koehn and Hoang (2007) and Bojar (2007). It used the lemmas and the output of the morphological analysis to decompose the translation process and use separate components to handle the transfer of lexical and grammatical information. In order to achieve this, the baseline system was extended with additional translation tables mapping SL lemmas to TL lemmas and SL morphology tags to TL morphology tags, respectively. In the target language, a generation model was used to transform lemmas and morphology tags into word forms. The results reported by Koehn and Hoang (2007) strongly indicate that this translation approach is not sufficient on its own; instead, the decomposed translation approach should be combined with a standard word form translation model so that one can be used in those cases where the other fails. This configuration was therefore adopted for our experiments.

The analytical translation approach fails to achieve any significant score improvement with the full parallel corpus. Closer examination of the MT output reveals that the strategy of using lemmas and morphological information to translate unknown word forms works in principle, as shown by the following example:

Input: Molly har visat mig bröllopsfotona.
Reference: Molly har vist mig fotoene fra brylluppet.
Baseline: Molly har vist mig bröllopsfotona.
Analytical: Molly har vist mig bryllupsbillederne.

In this sentence, there can be no doubt that the output produced by the analytical system is superior to that of the baseline system. Where the baseline system copied the Swedish word *bröllopsfotona* 'wedding photos' literally into the Danish text, the translation found by the analytical model, *bryllupsbillederne* 'wedding pictures', is both semantically and syntactically flawless. Unfortunately, the reference translation uses different words, so the evaluation scores will not reflect this improvement.

The lack of success of analytical translation in terms of evaluation scores can be ascribed to at least three factors: Firstly, there are relatively few vocabulary gaps in our data, which is due to the size of training corpus. Only 1.19 % (1,311 of 109,823) of the input tokens are tagged as unknown by the decoder in the baseline system. As a result, there is not much room for improvement with an approach specifically designed to handle vocabulary coverage, especially if this approach itself fails in some of the cases missed by the baseline system: Analytical translation brings this figure down to 0.88 % (970 tokens), but no further. Secondly, employing generation tables trained on the same corpus as the translation tables used by the system limits the attainable gains from the outset, since a required word form that is not found in the translation table is likely to be missing from the generation table, too. Thirdly, in case of vocabulary gaps in the translation tables, chances are that the system will not be able to produce the optimal translation for the input sentence. Instead, an approach like analytical translation aims

to find the best translation that can be derived from the available models, which is certainly a reasonable thing to do. However, when only one reference translation is used, current evaluation methods will not allow alternative solutions, uniformly penalising all deviating translations instead. While using more reference translations could potentially alleviate this problem, multiple references are expensive to produce and just not available in many situations. Consequently, there is a systematic bias against the kind of solutions analytical translation can provide: Often, the evaluation method will assign the same scores to untranslated gibberish as to valid attempts at translating an unknown word with the best means available.

6 Experiments with reduced corpora

We tested SMT systems trained on reduced corpora in two experimental conditions. In the *symmetric* condition, the systems described in the previous section were trained on a parallel corpus of 9,000 subtitles, or around 100,000 tokens per language, only. This made it possible to study the behaviour of the systems with little data. In the *asymmetric* condition, the small 9,000 subtitle parallel corpus was used to train the translation models and lexical reordering models. The generation and language models, which only rely on monolingual data in the target language, were trained on the full 900,000 subtitle dataset in this condition. This setup simulates a situation in which it is difficult to find parallel data for a certain language pair, but monolingual data in the target language can be more easily obtained. This is not unlikely when translating from a language with few electronic resources into a language like English, for which large amounts of corpus data are readily available.

The results of the experiments with reduced corpora follow a more interesting pattern. First of all, it should be noted that the experiments in the *asymmetric* condition consistently outperformed those in the *symmetric* condition. Evidently, Statistical Machine Translation benefits from additional data, even if it is only available in the target language.

The training corpus of 9,000 segments or 100,000 tokens per language used in the *symmetric* experiments is extremely small for SMT; in comparison to the training sets used in most other studies, this set is tiny. Consequently, one would expect the translation quality to be severely impaired by data sparseness issues, making it difficult for the Machine Translation system to handle unseen data. This prediction is supported by the experiments: The scores are improved by all extensions that allow the model to deal with more abstract representations of the data and thus to generalise more easily. The highest gains in terms of BLEU and NIST scores result from the morphology language model, which helps to ensure that the TL sentences produced by the system are well-formed.

Interestingly enough, the relative performance of the lexical reordering models runs contrary to the findings obtained with the full corpus. Lexical reordering models turn out to be helpful when conditioned on lemmas or POS tags, whereas lexical reordering conditioned on word forms neither helps nor hurts. This is probably due to the fact that it is more difficult to gather satisfactory information about reordering from the small corpus. The reordering probabilities can be estimated more reliably after abstracting to lemmas or POS tags.

In the *asymmetric* condition, the same phrase tables and lexical reorderings as in the *symmetric* condition were used, but the generation tables and language models were trained on a TL corpus 100 times as large. The benefit of this larger corpus is obvious already in the baseline experiment, which is completely identical to the baseline experiment of the *symmetric* condition except for the language model. Clearly, using additional monolingual TL data for language modelling is an easy and effective way to improve an SMT system.

Furthermore, the availability of a larger data set on the TL side brings about profound changes in the relative performance of the individual systems with respect to each other. The POS language model, which proved useful in the *symmetric* condition, is detrimental now. The morphology language model does improve the BLEU score, but only by a very small amount, and the effect on the NIST score is slightly negative. This indicates that the language model operating on word forms is superior to the abstract models when it is trained on sufficient data. Likewise, all three lexical reordering models hurt performance in the presence of a strong word form language model. Apparently, when the language model is good, nothing can be gained by having a doubtful reordering model

trained on insufficient data compete against it.

The most striking result in the *asymmetric* condition, however, is the score of the analytical translation model, which achieved an improvement of impressive 1.9 percentage points in the BLEU score along with an equally noticeable increase of the NIST score. In the *asymmetric* setup, where the generation model has much better vocabulary coverage than the phrase tables, analytical translation realises its full potential and enables the SMT system to produce word forms it could not otherwise have found.

In sum, enlarging the size of the target language corpus resulted in a gain of 2.7 percentage points BLEU on the baseline score of the *symmetric* condition, which is entirely due to the better language model on word forms and can be realised without linguistic analysis of the input. By integrating morphological analysis and lemmas for both the SL and the TL part of the corpus, the leverage of the additional data can be increased even further by analytical translation, realising another improvement of 1.9 percentage points, totalling 4.6 percentage points over the initial baseline.

7 Conclusion

Subject to a set of peculiar practical constraints, the text genre of film subtitles is characterised by short sentences with a comparatively simple structure and frequent reuse of similar expressions. Moreover, film subtitles are a text genre designed for translation; they are translated between many different languages in huge numbers. Their structural properties and the availability of large amounts of data make them ideal for Statistical Machine Translation. The present report investigates the potential of incorporating information from linguistic analysis into the Swedish–Danish phrase-based SMT system for film subtitles presented by Volk and Harder (2007). It is based on a subset of the data used by Volk and Harder, which has been extended with linguistic annotations in the Constraint Grammar framework produced by the DanGram parser (Bick, 2001). We integrated the annotations into the SMT system using the factored approach to SMT (Koehn and Hoang, 2007) as offered by the Moses decoder (Koehn et al., 2007) and explored the opportunities offered by factored SMT with a number of experiments, each adding a single additional component into the system.

When a large training corpus of around 900,000 subtitles or 10 million tokens per language was used, the gains from adding linguistic information were generally small. Minor improvements were observed when using additional language models operating on part-of-speech tags and tags from morphological analysis. A technique called analytical translation, which enables the SMT system to back off to separate translation of lemmas and morphological tags when the main phrase table does not provide a satisfactory translation, afforded slightly improved vocabulary coverage. Lexical reordering conditioned on word forms also brought about a minor improvement, whereas conditioning lexical reordering on more abstract categories such as lemmas or POS tags had a detrimental effect.

On the whole, none of the gains was large enough to justify the cost and effort of producing the annotations. Moreover, there was a clear tendency for complex models to have a negative effect when the information employed was not selected carefully enough. When the corpus is large and its quality good, there is a danger of obstructing the statistical model from taking full advantage of the data by imposing clumsily chosen linguistic categories. Given sufficient data, enforcing manually selected categories which may not be fully appropriate for the task in question is not a promising approach. Better results could possibly be obtained if abstract categories specifically optimised for the task of modelling distributional characteristics of words were statistically induced from the corpus.

The situation is different when the corpus is small. In a series of experiments with a corpus size of only 9,000 subtitles or 100,000 tokens per language, different manners of integrating linguistic information were consistently found to be beneficial, even though the improvements obtained were small. When the corpus is not large enough to afford reliable parameter estimates for the statistical models, adding abstract data with richer statistics stands to improve the behaviour of the system. Compared to the system trained on the full corpus, the effects involve a trade-off between the reliability and usefulness of the statistical estimates and of the linguistically motivated annotation, respectively; the difference in the results stems from the fact that the quality of the statistical models strongly depends on the amount of data

available, whilst the quality of the linguistic annotation is about the same regardless of corpus size. The close relationship of Swedish and Danish may also have impact: For language pairs with greater grammatical differences, the critical corpus size at which the linguistic annotations we worked with stop being useful may be larger.

Our most encouraging findings come from experiments in an asymmetric setting, where a very small SL corpus (9,000 subtitles) was combined with a much larger TL corpus (900,000 subtitles). A considerable improvement to the score was realised just by adding a language model trained on the larger corpus, which does not yet involve any linguistic annotations. With the help of analytical translation, however, the annotations could be successfully exploited to yield a further gain of almost 2 percentage points in the BLEU score. Unlike the somewhat dubious improvements in the other two conditions, this is clearly worth the effort, and it demonstrates that factored Statistical Machine Translation can be successfully used to improve translation quality by integrating additional monolingual data with linguistic annotations into an SMT system.

References

Daniel Becquemont. 1996. Le sous-titrage cinématographique : contraintes, sens, servitudes. In Yves Gambier, editor, *Les transferts linguistiques dans les médias audiovisuels*, pages 145–155. Presses universitaires du Septentrion, Villeneuve d'Ascq.

Eckhard Bick. 2001. En Constraint Grammar parser for dansk. In *8. Møde om udforskningen af dansk sprog*, pages 40–50, Århus.

Ondřej Bojar. 2007. English-to-Czech factored Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 232–239, Prague.

Paul R. Cohen. 1995. *Empirical methods for Artificial Intelligence*. MIT Press, Cambridge (Mass.).

Jorge Díaz-Cintas and Aline Remael. 2007. *Audiovisual Translation: Subtitling*, volume 11 of *Translation Practices Explained*. St. Jerome Publishing, Manchester.

George Doddington. 2002. Automatic evaluation of machine translation quality using *n*-gram co-occurrence statistics. In *Proceedings of the second International conference on Human Language Technology Research*, pages 138–145, San Diego.

Fred Karlsson. 1990. Constraint Grammar as a framework for parsing running text. In *COLING-90. Papers presented to the 13th International conference on Computational Linguistics*, pages 168–173, Helsinki.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Conference on empirical methods in Natural Language Processing*, pages 868–876, Prague.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Edmonton.

Philipp Koehn, Amittai Axelrod, et al. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *International workshop on spoken language translation*, Pittsburgh.

Philipp Koehn, Hieu Hoang, et al. 2007. Moses: open source toolkit for statistical machine translation. In *Annual meeting of the Association for Computational Linguistics: Demonstration session*, pages 177–180, Prague.

Franz Josef Och. 2003. Minimum error rate training in Statistical Machine Translation. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo (Japan).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia. ACL.

Jan Pedersen. 2007. *Scandinavian subtitles. A comparative study of subtitling norms in Sweden and Denmark with a focus on extralinguistic cultural references*. Ph.D. thesis, Stockholm University, Department of English.

Andreas Stolcke. 2002. SRILM: an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, Denver (Colorado).

Teresa Tomaszkiewicz. 1993. *Les opérations linguistiques qui sous-tendent le processus de sous-titrage des films*. Wydawnictwo Naukowe UAM, Poznań.

Martin Volk and Søren Harder. 2007. Evaluating MT with translations or translators. What is the difference? In *Proceedings of MT Summit XI*, pages 499–506, Copenhagen.

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206